Working paper

2019-02

Statistics and Econometrics
ISSN 2387-0303

Social Pressure or Rational Reactions to Incentives? A Historical Analysis of Reasons for Referee Bias in the Spanish Football

Stefano Cabras, J. James Reade and J.D. Tena

uc3m | Universidad Carlos III de Madrid

Social Pressure or Rational Reactions to Incentives? A Historical Analysis of Reasons for Referee Bias in the Spanish Football.

By

Stefano Cabras **Universidad Carlos III de Madrid (Spain) and Università degli Studi di Cagliari (Italy)**

J. James Reade **University of Reading (UK)**

J.D. Tena[1] **University of Liverpool (UK) and Università degli Studi di Sassari (Italy)**

Abstract

A relevant question in social science is whether cognitive bias can be instigated by social pressure or is it just a rational reaction to incentives in place. Sport, and association football in particular, offers settings in which to gain insights into this question. In this paper we estimate the determinants of the length of time between referee appointments in Spanish soccer as a function of referee decisions in favour of the home and away team in the most recent match by means of a deep-learning model. This approach allows us to capture all interactions among a high-dimensional set of variables without the necessity of specifying them beforehand. Furthermore, deep-learning models are nowadays the state of the art among the predicting models which are needed and here used for estimating effects of a cause. We do not find strong evidence of an incentive scheme that counteracts well-known home referee biases. Our results also suggest that referees are incentivised to deliver a moderate amount of surprise in the outcome of the game what is consistent with the objective function of consumers and tournament organisers.

Keywords: Causal analysis; Deep-learning model; Referee bias; Social pressure; Sport.

1. Introduction

[1] Corresponding author: J.D. Tena, University of Liverpool Management School, Chatham Street, L69 7ZH, email: jtena@liverpool.ac.uk.

Subjective decisions play an important role in many domains of life such as, for example, family, clinical judgement, jurisdiction, management, and policy. Statistics in social science has typically analysed the different problems associated to this issue under, at least, two different perspectives. First, a branch of literature has explored the importance of incentives to explain different types of human decisions regarding, for example, providing information in surveys (Stecklov et al, 2018), and legislative behaviour (Titiunik and Feher, 2017).

A second approach considers the concept of cognitive bias introduced by Tversky and Kahneman (1974) as a mistake in reasoning as a result of using heuristic rules to reduce the complexity of the problem. These rules are affected by subjective preferences or beliefs and therefore, decisions are not explained in terms of monetary incentives but from the inability of the decision maker to deal with a complex amount of information in a limited amount of time typically in situations of high social pressure.

Sports economics, and football in particular, has offered many natural experiment situations for applied statisticians, psychologists and behavioural economists to test for the presence of cognitive bias in the process on making decisions under external pressure; see for example Garicano et al. (2005) and Buraimo et al. (2010) just to mention two examples. There are important reasons for that. First, data on referee decisions are recorded and scrutinised. Second, football is an interesting sport in this respect as there is a high degree of discretion in many referee decisions such as the added time at the end of the game and the intentionality or otherwise of a handball and a violent action. Moreover, football is a low scoring game and any decision can potentially have an important impact on the final score of the match.

However, as far as we are concerned, with the only exception of Price et al. (2012), the previous sport literature on referee cognitive bias has ignored the incentive scheme that football referees face when making their decisions. This is a very relevant issue as to judge a decision as irrational or biased is necessary to explore whether it is just a rational response to the incentive scheme offered by the agent. Price et al. (2012) hypothesised that referee resolutions could aim to increase consumer satisfaction by being favourable to home teams and trailing teams in a set of play-off basketball games. Moreover, they provided indirect empirical evidence for this hypothesis.

In this paper we explore the incentive scheme of referees in the top tier of the Spanish football league (Primera División) by estimating how a referee decision in a match could affect the number of weeks that a referee must wait to be appointed for the next game. Of course, being able to referee again in a short period of time is not the only reward that a referee can obtain from his work, but it is the only one that can be consistently and publicly observed through all

our analysis period. Traditionally, 'la nevera', or 'the fridge' in English, is an expression in the Spanish jargon to apply to a referee who is punished by not working for several weeks because of important decision mistakes in his last match. The institution responsible to make this decision, the Spanish Football Federation, does not report information about which referees are in this situation but its existence itself is an empirical question to be investigated.

We investigate how decisions on penalties and the number of sent off players due to yellow and red cards affect the length of time for the next referee appointment. Moreover, we study if their consequences are significantly different when they favour the home and the away team. This is relevant to get evidence on whether the incentive scheme offered to soccer referees counteracts or explain the biased reported in the previous literature. We also analyse whether referees are incentivised to deliver expected results. In principle, it is reasonable to assume that organisers promote some amount of surprise that adds excitement to the competition but at the same time, they do not want a referee with a very large influence in the outcome of the match.

Our causal analysis is conducted by means of a deep-learning (DL) model, see Schmidhuber (2015). Causality is here understood as the estimation of effects of causes. This requires accounting for main confounders along with a flexible and reliable prediction model for the response variable (the DL model). We took this approach because of two important reasons. First, a DL model is a sophisticated approach that enriches our analysis by allowing us to investigate the contribution of an extremely large number of variables in the model and the estimation of highly non-linear causal responses of our focus variables that can be potentially interacting with many covariates without the necessity of specifying them beforehand and also performing model selection procedures. This is possible because of the availability of a large dataset which allows for model selection (which is intrinsic in DL) as long as estimation of effects. More importantly, under most standard econometric approaches, model specification and the results of analysis themselves can be influenced by researcher cognitive biases. Silberzahn and Uhlmann (2015) report the results of a crowdsourcing analysis where different researchers were supplied with the same dataset asking them to provide an empirical estimation for a specific answer on racial bias for football referees. They found substantial differences in their responses. DL specifications are not subjective but decided by the machine learning algorithm. Essentially, the DL can be viewed as a complicated regression model in which we do not specify principal effects along with their interactions, but allowing for all possible effects (subject to the specified regressors) as data will estimate the most appropriate model in a non-parametric fashion. This is what avoids analyst subjectiveness.

The remaining of this paper is structured as follows. Next section discussed the related literature on soccer referee bias. In the following section, we present our data and the empirical

approach used in the analysis. Estimation results are shown and discussed in Section 4 and some concluding remarks follow in Section 5.

2. Related Literature

Garicano et al. (2005) initiated the analysis on football referee bias by studying the tendency for referees in the Spanish football league to increase stoppage time in close games when the home team is trailing compared to a situation when the home team is leading. More specifically, when the home team is trailing by one goal, the average added-on time increases by 35% above the norm; but when the home team is leading by one goal, there is a 29% reduction in stoppage time compared to the average. Garicano et al. (2005) inspired other studies to extend this analysis to other leagues. Thus, Sutter and Kocher (2004) using information from the German Bundesliga during the 2001 season found that if the home team is ahead by one goal or the scores are level, added-on time is between 20 to 50 seconds lower than when the home team is trailing by one goal.

The literature has also found a lower evidence of disciplinary sanctions (in terms of red and yellow cards) for home teams in the English Premier League (Dawson et al., 2007), the top tier of the Bundesliga and the English Premier League (Buraimo et al., 2010) and in European cup matches (Dawson and Dobson, 2010).

However, the analysis of referee biases is not restricted to home advantage. Although it is out of the scope of this paper a comprehensive revision of the literature, we mention three examples; the first two use data from the American National Basketball Association League, and are Price and Wolfers (2010) and Price et al. (2012). Price and Wolfers (2010) find evidence of referee preferences for players whose ethnicity is the same as the majority of the referee crew while results in Price et al. (2012) explore other types of biases such as referee predilection for close games and loser teams. The third example, Gallo *et al* (2013) consider implicit discrimination against black African players in the English Premier League via the incidence of disciplinary measures.

A main reason to explain the presence of referee bias is the role that social pressure exerts on their decisions. In the case of home advantage, pressure can be a function of attendance. For example, Garicano et al. (2005) and Pettersson-Lidbom and Priks (2010) find that a significant amount of home bias in the top tier of the Spanish and the Italian League respectively are influenced by the ratio of attendance to stadium capacity but Buraimo et al. (2010) do not find evidence to support the hypothesis that a lower incidence of disciplinary actions for the home team is explained by the size of the crowd but it can be explained by the absence of running

tracks in stadia.

However, pressure not necessarily comes from attendance. In fact, social attention can also affect the decision-making process. Pope et al. (2018) replicate the analysis on racial bias by Price and Wolfers (2010) using more recent data finding that the effects are no longer significant when they consider the 2007-2010 period. Given that the NBA does not appear to have made any attempt to address the issue, the authors' interpretation is that increased awareness of racial discrimination in NBA refereeing was sufficient to eliminate that racial discrimination. An additional explanation, which is particularly relevant in this paper, for the presence of home bias is provided by Price et al. (2012). They hypothesised that referee preferences for the home team, rather than an irrational decision, could serve to increase consumer satisfaction. A similar situation occurs with close games. They empirically tested this idea by estimating the impact from previous estimates of referee preferences for home teams, close games and differences in winning percentages between home and away teams (Match-up Coefficients) on the probability that a referee is assigned to a playoff game which can be considered as an obvious and visible form of compensation. Only Match-up Coefficients turns out to be significant in that regression. This was interpreted as an indirect evidence on the existence of incentives for bias.

3. Data and Empirical Strategy

3.1. Data

The Spanish Primera Division is the top tier of the Spanish football. The first edition of the two top tiers Spanish leagues took place in season 1928-29 and the tournament was played every year since then with the only exception of the period between 1936 and 1939 due to the Spanish Civil War. The competition has worked as a round-robin tournament where clubs are promoted and relegated based on performance. Throughout the history of the competition, only three clubs have been present in all of the editions of the Spanish Primera Division: Real Madrid, Atletico de Bilbao and FC Barcelona.

The Real Federacion Española de Futbol (RFEF) is the organiser of the Spanish Football League and its subordinate division, Comite Tecnico de Arbitros (CTA)[2] is the responsible to appoint referees to the different games. Through the history, there has been 31 different presidents of the RFEF. The last one, Jose Maria Villar, has been the longest-lasting president being in charge from 1988 to the end of our analysis period. The way to allocate referees has been affected by different policy stance periods but, with the exceptions of seasons 1953/54-1956/7, 1971/72 to 1975/76 and 1996/7 to 2004/05 were referees were randomly appointed, there has been some degree of discretion in these decisions. We collected match level data for

[2] Also known as Comite Nacional de Arbitros

the whole history (from 1929 to 2017) of the top tier of the Spanish League from the database BDFUTBOL at the url: https://www.bdfutbol.com. For each game, the variable whose response we want to analyse is the number of weeks a football referee must wait until he referees the next football match (*time*). We are interested in cases with large number of weeks as small amounts of time are supposed to be purely random. Therefore, we consider cases with more than two weeks that a referee must wait. Two important features must be mentioned about this variable. Firstly, it is measured in terms of game weeks, the weeks when games take place, rather than actual calendar weeks. Secondly, this variable has a number of missing values which represent around 2% of the sample. This is due to referees being demoted, or retiring. Despite this, we have a total of 19636 observations for 22 variables, some of which are categorical with many also levels (the referee identity variable has 661 levels referring to an equal number of referees). This corresponds to a design matrix with 1152 columns, which is so large that any subjective variable selection is prohibitive.

The covariates included in the model are the number of sent off players with two yellow cards for the home and away teams, *home2yellow* and *away2yellow* respectively; similar variables are defined for the number of sent off players with a red card, *homered* and *awayred*; the number of penalties in favour of the home and away teams, *homepen* and *awayred*. We also consider a few dummy variables to indicate, for example, the home and away team, the referee, the number of scored goals for the home and for the away team, outcome of the game, Villar period, round and season. We also consider the Brier Score of the match. This was obtained by using the Elo ratings of the teams to specify ordered probit models estimated with a window of 5 seasons. This model was used to obtain probabilities of home victories, draws and away victories that were considered to compute the Brier Score of the match.

The following table shows a descriptive statistic for these variables.

Table 1. *Descriptive Statistics*

Variable	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
round	19,636	18.498	10.503	1	10	27	44
homeyellow	19,636	1.811	1.754	0	0	3	11
home2yellow	19,636	0.057	0.248	0	0	0	3
homered	19,636	0.065	0.262	0	0	0	4
homepen	19,636	0.129	0.355	0	0	0	3
awayyellow	19,636	2.061	1.861	0	0	3	12
away2yellow	19,636	0.075	0.279	0	0	0	3
awayred	19,636	0.080	0.293	0	0	0	5
awaypen	19,636	0.074	0.270	0	0	0	2
homegoals	19,636	1.648	1.398	0	1	2	12
awaygoals	19,636	0.977	1.039	0	0	2	8

briar.score	19,636	0.111	0.120	0	0	0.2	1
time	19,636	4.021	2.618	3	3	4	44
Factors variables		# of levels		Levels –(Frequency)			
Season	19,636	87					
Referee IF	19,636	661					
Outcome	19,636	3	0 (4273)	0.5 (5039)	1 (10324)		
TeamID - Home	19,636	231					
TeamID – Away	19,636	236					
Villar	19,636	2	TRUE (11774)	FALSE (7862)			

3.2. Rationality behind causal estimation of referee relegation.

The aim of this study is to estimate the causal effect for a referee in previous game i with respect to its action D_i on the time he has to wait for refereeing the next match Y_i, such that the Average Treatment Effect on for a referee in game i is defined as $ATE_i = \mathbf{E}_{\pi(Y_i|X_i=x_i,Data)}(y_i(D_i = d_o) - Y_i(D_i = d_c)|X_i = x_i)$, where $D_i = d_o$ is the observed action in previous game i and $D_i = d_c$ represent the action he could have taken. For instance, $D_i = d_o$ could be that we have observed zero red cards, while $D_i = d_c$ represent the *what if* situation, e. g. what if on previous match i, the referee would have shown two red cards. The latter represent the counterfactual situation which corresponds an estimation of the counterfactual wait time Y_i considered the random variable that must be predicted as neither $D_i = d_c$ have been observed, nor Y_i. Furthermore, ATE is defined upon the expectation of the random variable Y_i conditional on all information on previous game i including that on the referee itself (i. e. its ID) and the observed sample. The predictive model $Y|X, D$ (that for sake of simplicity, we refer to it as $Y|X$, intending that X includes D) we employed and detailed in the next section, is an approximation of the Bayesian predictive distribution (the one that appears on the sub index of the expectation). This approach is alternative to one that matches referees with, say, 0 red cards on a match with referees with 2 red cards based on their estimated propensity scores. The application of the propensity score methodology is only possible if there is a region of common support between the two groups of referees. In our problem such common region also includes, for instance the teams involved in the match along with whether they were home or away. Such a common region simply does not exist. Moreover, in order to satisfy the strong ignorability assumption, required for causal inference, we have to account for all possible collected confounding variables (i. e. elements of vector X_i) along with their interaction in the predictive model for Y_i. The dimension of X_i is such that it is impractical to

specify which covariables use and their interaction beforehand, but let the method estimate the most suitable model for predicting Y_i. This is one of the main advantages of the methodology considered in this paper as it allows for the identification and estimation of different types of nonlinear interactions between the treatment variables and the different covariates without the necessity of estimating different models for each interaction. Finally, in order to draw causal conclusions, it is necessary to have an almost perfect prediction model to capture all relations between the response and the predicting variables including that on which we want to evaluate the causal effect. Again, the deep-learning predictive model that we are using represents the actual state of the art in model prediction with high dimensional data and large datasets.

3.3. The deep-learning predicting model

A deep-learning (DL) model is a neural network with many layers of neurons (Schmidhuber 2015). DL refers mostly to an algorithmic approach rather than a specific probabilistic model, although both components are present in DL (see Breiman, 2001, for the merits of including both elements). Each neuron is a deterministic function such that two connected neurons correspond to a function of a function along with an associated weight w. Essentially, for a response variable Y_i for referee i and a predictor variable X_i (or an entry of the design matrix X) we have to estimate $Y_i = w_1 f_1\left(w_2 f_2\left(\cdots\left(w_k f_k(X_i)\right)\right)\right)$, and the larger the k is the more the network is "deep". With many stacked layers of neurons all connected (a.k.a. dense layers) it is possible to capture high non-linearities and all interactions among variables. The approach to model estimation underpinned by a DL model is that of compositional function against that of additive function underpinned by the usual regression techniques including the most modern ones (i.e. $Y_i = w_1 f_1 + w_2 f_2 + \cdots + w_k f_k(X_i)$). See Schmidhuber (2015) for more details.

The DL model (in this case a non recurrent neural network) can be also interpreted, for the set of observations denoted by $Data$, as a posterior mode estimation of $Pr(Y|X, Data)$ (Polson and Sokolov, 2017) from gaussian process priors through its probabilistic nature, which ultimately gives a strong statistical support to the analysis conducted here. However, due to its complexity, the whole distribution $Pr(Y|X, Data)$ cannot be evaluated but only its mode. This prevents a full Bayesian analysis of the problem, but it implies that the causal effects estimated here are those which maximise the probabilistic density given the observed data.

In this setting Y_i and X_i can be scalar or vector and in particular Y is the scalar random variable of times (in weeks) and X is vector of dimension 1152 as it incorporates the above predictors with all their levels when they are factors (i.e. there are 661 referees and hence 661 dummy variables representing the effect of the referee.).

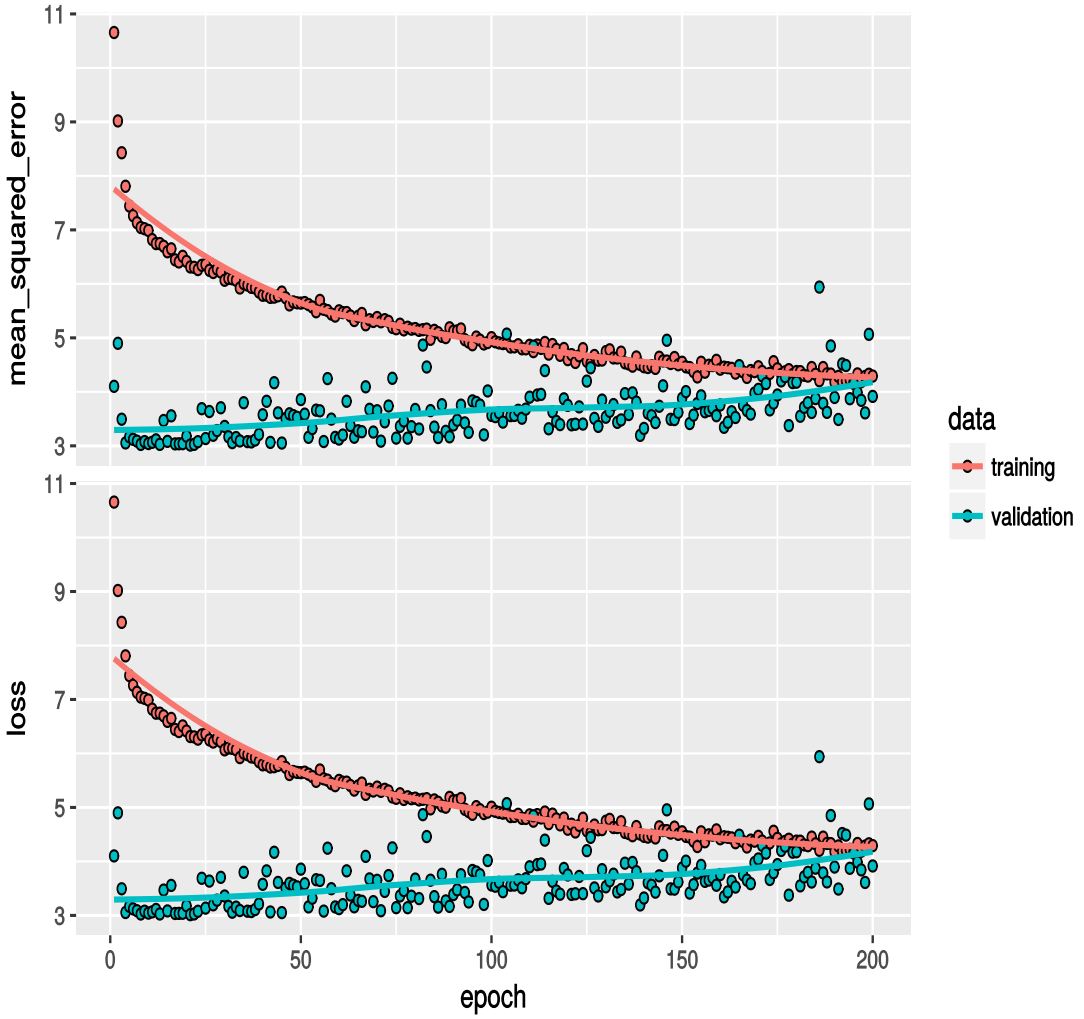
Estimating a DL model consists of estimating the vectors $w_1, \cdots, w_k$. Estimation requires the evaluation of a multidimentional gradient which cannot be evaluated jointly for all observations because of its dimensionality and complexity. Recalling that the derivative of a composite function is defined as the product of the derivative of inner functions (i.e. the usual chain rule $(f \circ g)' = (f' \circ g) \cdot g'(\quad)$) which is implemented for purposes of computational feasibility as a tensor product. Such tensor product is evaluated independently for batch of observations and it is implemented in the open source software Google Tensor Flow (Abadi et al. 2015) running on a NVidia Quadro GPU. Tensor product, independent evaluation, low cost processors (per unit) as GPU makes the DL approach popular nowadays. There are different optimisation algorithms to estimate w_s and we used the Adaptive Subgradient Methods (ADAGRAD) (Duchi et al, 2011) in order to minimize the squared loss function, i.e. w are estimated in order to minimize $\sum_{i=1}^{N=19636}(y_i - \hat{y}_i)^2$ the quadratic differences between Y_i and the prediction $\hat{Y}_i = \hat{w}_1 f_1\big(\hat{w}_2 f_2(\cdots \hat{w}_k f_k(X_i)\,)\big)$.

The structure of the model is the following: we have two dense layers, separated by a normalization batch layer and a dropout layer at 40% to avoid overfitting and achieve model parsimony. We have around 60 thousand parameters (i.e. weights) to be updated. Of course, some weights will be zero as they do not contribute to the gradient of the quadratic loss function and this avoids overfitting and implement the variable selection needed with 1152 predictors. Furthermore, to achieve stability in estimation we introduced a normalization batch between the two sets of hidden layers (Ioffe and Szegedy 2015). Normalization batch is the usual operation of variable standardization (i.e. mean zero and variance one) applied to weights connecting two sets (layers) of all connected neurons. Ioffe and Szegedy (2015) show that this

operation allows for better stability in the gradient of the whole function $Y|X$ estimated with the DL model.

The following graphs show the result of the optimization procedure, iterated for 200 hundred steps. The loss in the training set (a sample subset randomly defined at a given step and used in the gradient) is practically monotone decreasing meaning that the model is learning from the data. On the other hand, the loss in the validation set (a subset of the training sample not used for fitting at that particular epoch (optimization step)) is almost always below that in the training set (used to calculate the weights) indicating that the model does not overfit the data (remember: there are more parameters than observations).

Figure 1. *Results of the Optimization Procedure*



The estimated model can predict 50% of the variability of the response variable. This is indeed a significant result as it indicates that the length of time between referee appointments is not purely random (as expected) but can be forecasted in some way by a DL model

4. Analysis

In this section, we explore the causal effects of referee decisions on the length of time that a referee must wait for his next appointment. More specifically, we estimate the impact of disciplinary decisions in terms of yellow and red cards, penalties and an indicator of how surprising the outcome of the last game was, which is measured by Brier Score.

Three main hypotheses will be tested informally by means of the analysis of the prediction uncertainty arising from the DL model. The first one concerns the preference of the organisers for the home or away team. Although we cannot measure whether disciplinary decisions and penalties were fair or not, however, it is well known, from the literature, that there is consistent evidence (in particular for the Spanish League) of home referee bias regarding these decisions;

see, for example, Garicano et al. (2005). Therefore, organisers trying to counteract these biases must penalise referee decisions in favour of the home team relatively more than similar decisions in favour of the away team. However, on the other hand, organisers themselves could be also affected by the pressure generated by home supporters that could bias their decisions. Thus, there is no clear hypothesis on whether the Spanish Federation is going to punish more or fewer decisions in favour of the home compared to the away team.

Another interesting hypothesis to test concerns the evaluation of the cost of making decisions. Under the omission bias hypothesis (Samuelson and Zeckhauser, 1988), referees would have incentives to not making decisions as mistakes in actions are more obvious than mistakes in inactions. According to this, not awarding penalties or sending players off would be more profitable for the referee than making any decision in this respect.

A final issue of concern relates to the incentives that referees can face to deliver an unexpected result. Our hypothesis is that the Spanish Federation could incentivise a certain amount of surprise in the outcome of the games as this maintains interest in the competition.

DL models are non-parametric techniques which not provide information on the effect of covariates. However, as discussed in the previous section, it is possible to estimate the causal impact of a given referee decision by comparing the expected response under factual and counterfactual observations. We do this by comparing the observed values of the response variable in a factual and the expected (i.e. the maximum a posteriori) in the counterfactual situation.

In particular, counterfactual estimation is obtained by the fitted DL model including all predictors with the intervention variable D changed in order to calculate the causal effect induced by a specific variable. This is founded on the fact that a model which perfectly predicts the response can be potentially used for causal inference. Formally, let $\tilde{\mathrm{X}}$ be the matrix of confounding variables and let D be the intervention variable representing the counterfactual situation, i.e. $\tilde{\mathrm{X}}$ does not have the intervention variable (this is why $\tilde{\mathrm{X}}$ and not just X). The range of factual values is set to be $\mathrm{D} = \{0,1,2,3,4\}$ while its assigned counterfactual values is $\mathrm{D} = \{4,3,2,1,0\}$. This implies that the resulting casual effects are estimated for variations in the counterfactual situation (in respect to the factual) of magnitudes $\mathrm{Z} = \{4,2,0,-2,-4\}$ in the intervention variable.

We evaluate casual effect for a given referee on the length of time to be appointed again of changing decisions about home2yellow, away2yellow, homered, awayred, homepen, awaypen and Brier Score. Given that our database corresponds to an extremely long historical period, a relevant question to answer is whether referees face different incentive schemes now and in the past. We estimate the causal effects before and after 1988 because it is the year that Jose Maria

Villar took over as president of the Spanish Football Federation, and remained in position until the end of our data period. This period is denoted with the name Villar. However, this distinction was not made in the case of home2yellow and away2yellow because yellow cards were only introduced in football after 1970.

We initiate our analysis by studying the expected penalization that a referee suffers as a consequence of decisions regarding the number or two yellow cards, red cards and penalties. These causal effects are shown in Figures 2 to 4. As our intervention variables are quantitative, causal effects are always represented by means of smoothing curves (which connects points on the horizontal axis). Such curves along with the 95% confidence intervals are obtained using GAM models (Wood, Pya, and Säfken 2016).

Figure 2. *Casual effect of variations in two yellow cards*

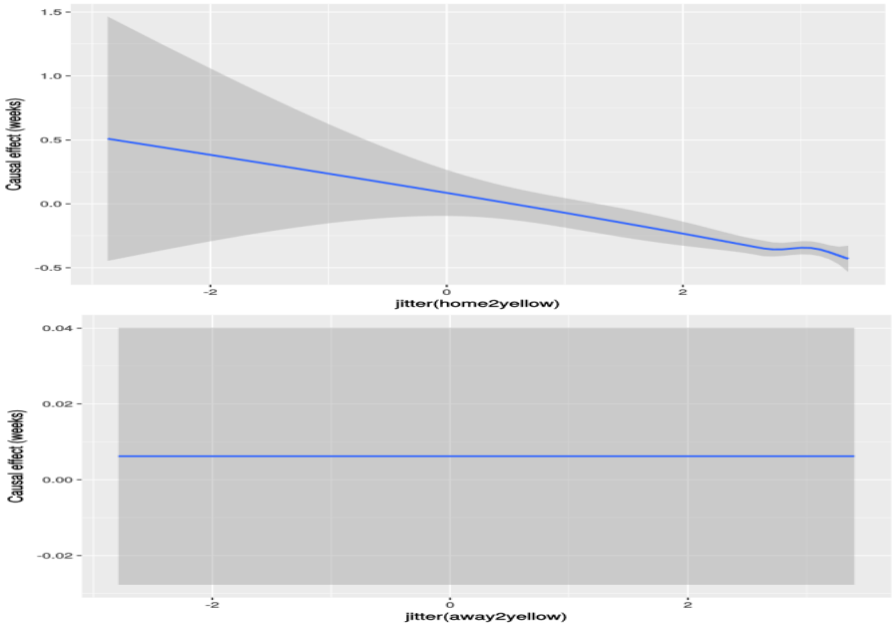


Figure 3. *Casual effect of variations in red cards*

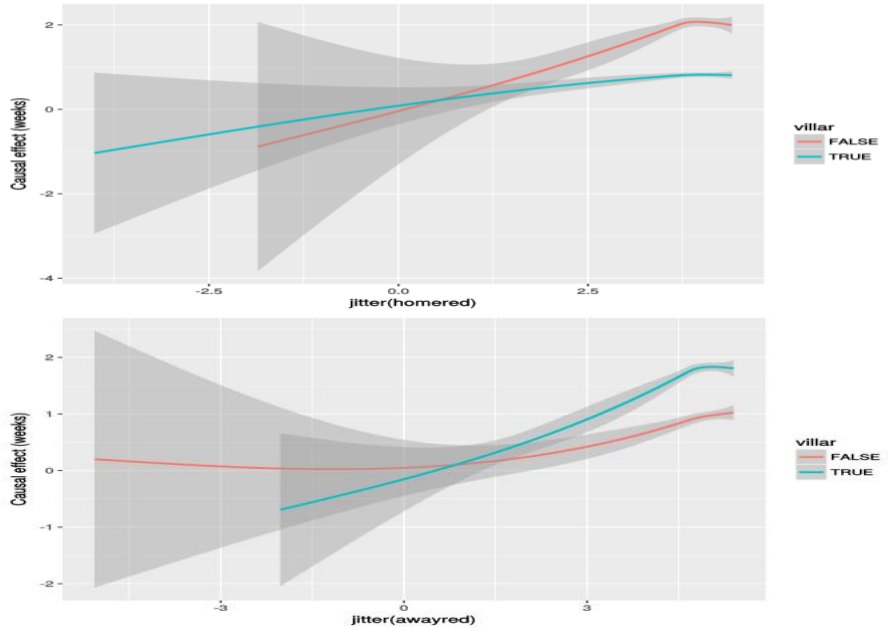
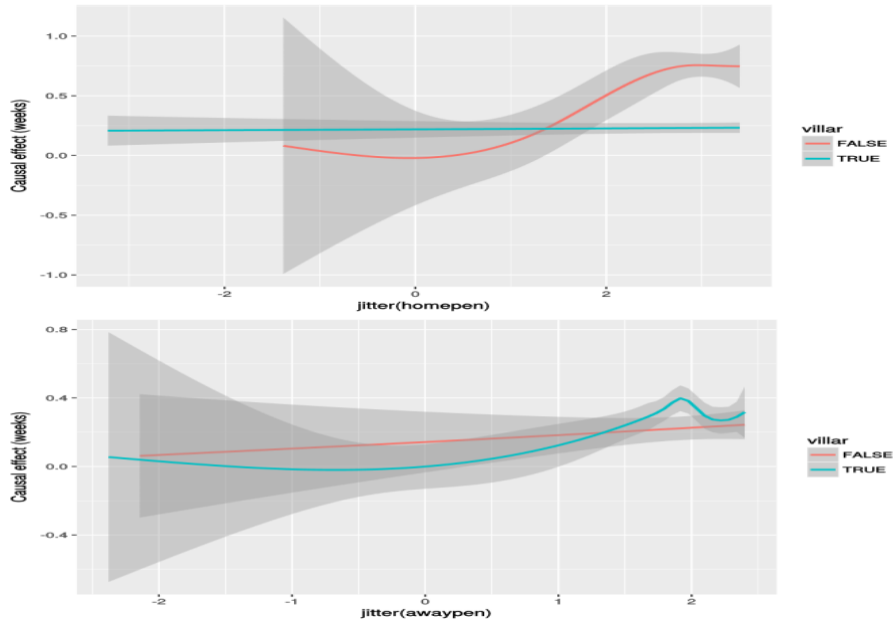


Figure 4. *Casual effect of variations in penalties*



A striking result is that increasing the number of two yellow cards for the home team reduces the number of weeks that a referee must wait to be appointed again while the opposite happens with red cards for the home team. A possible interpretation for this is that disciplinary sanctions by a referee are better understood when they are gradually taken rather than in an abrupt decision. However, we do not find evidence of any significant effect of variations in two yellow cards for the away team. This can be either due to the small number of observations for these events or to the fact that they are not significant. Remember that in order to avoid the influence of confoundings with respect to these intervention variables, personal history of the referee along with details of the match (teams, score, etc.) have been included in the analysis. Such confoundings are likely to better explain the response variable than the intervention

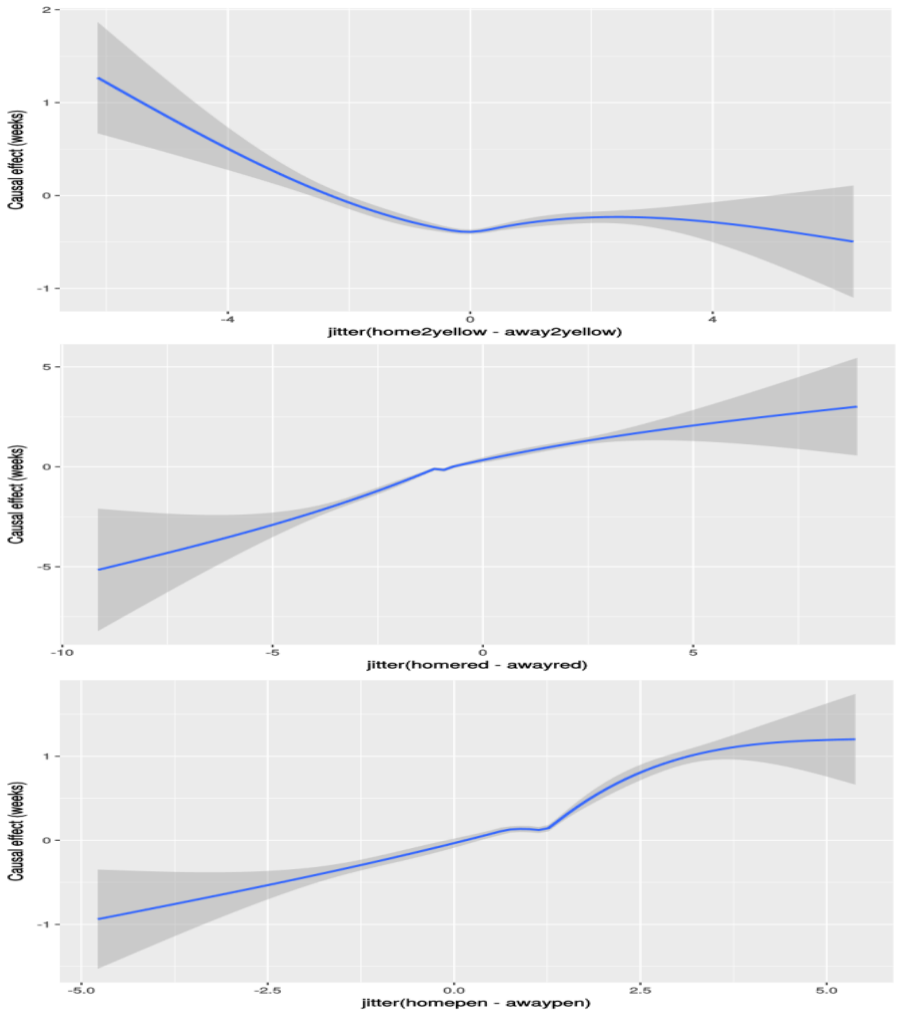
variable. Regarding variations in the number of penalties, an increase generates a penalisation both for the home and the away team. However, the punishment in terms of weeks for increasing the number of home penalties has been reduced in the most recent period compared to the past.

From figures 3 & 4, there is also evidence that both a high number of penalties and red cards increase the number of weeks that a referee must wait to be appointed again. The evidence also suggests the presence of an incentive scheme that favor inaction as there is no penalization for not whistling any penalty or not showing any red card.

When comparing the two different periods of analysis: before and after Villar took over as president of the Spanish Federation, it can be observed that in the Villar period there is a higher penalisation for a high number of away red cards while the opposite happens with the number of home red cards. However, when we turn our attention to penalty kicks, there is also evidence that the penalisation for a high number of penalties in favour of the home team has been reduced as well.

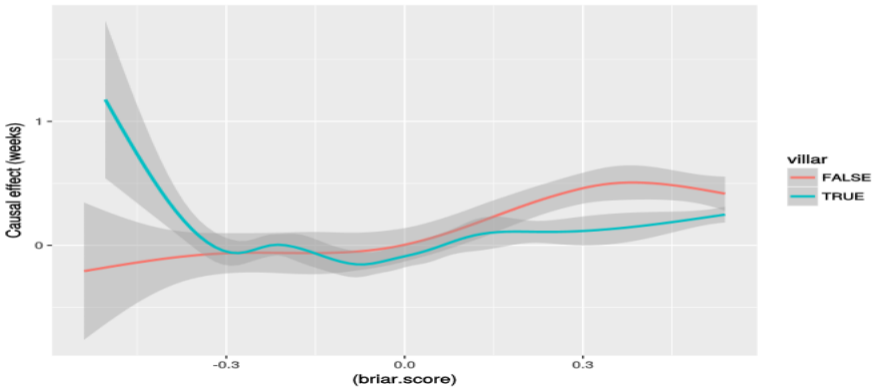
For a better analysis of the possible presence of incentives for home referee bias decisions, we estimate the causal effects of differences in 2 yellow cards, red cards, and penalties between the home and the away team for the most recent period. Figure 5 shows these estimations. Results indicate that referees have incentives to show relatively more red cards to the away team but more yellow cards to the home team. Regarding the causal impact of penalty kicks, our findings do not support the hypothesis of home bias incentives in terms of penalties as referees that increase the number of penalties in favour of the home team relatively to the away team have to wait for more time to be appointed again. Overall, we do not find definitive evidence of an incentive scheme, especially concerning red cards, to counteract the presence of home bias in referee decisions.

Figure 5. *Casual effects for home vs away referee decisions.*



Now we turn our attention to study how referees are penalised to deliver unexpected results. In order to study this we consider Brier Score that is an indicator of how surprising the outcome of a game is compared to what it was ex-ante expected. Causal effects for Brier Scores in the two reference periods are shown in Figure 6. There are many crossing points in the evolution of the causal effects in the two periods but they are significantly different for extremely low and high values of Brier Score suggesting that in the last period the Spanish Federation penalises referees who deliver highly expected or unexpected results. This is consistent with the insight that organisers care about the excitement of the competition and try to incentivise a moderate amount of surprise in the final score of the match. However, highly unexpected results generate a shock in social media which can call for the attention of organisers.

Figure 6 Casual effects for Brier Score



5. Concluding remarks

We have investigated the incentive scheme that referees in the top tier of the Spanish football league face when making different types of decisions. Our results indicate that referees are motivated to whistle a fewer number of penalties in favour of the home compared to the away team. However, there is evidence of incentives to referees to send off more away players compared to home players. A possible interpretation for this is that, compared to a red card, a penalty has a direct effect on the score of the match, and hence makes home bias more apparent. We have also found some evidence of incentives for referees to omit decisions regarding red cards and penalties and to deliver a moderate degree of surprise in the final outcome of the game.

The implementation of Video Assistant Referees (VAR) in the Spanish football from season 2018/19 will help referees reducing the amount of uncertainty they face when making decisions and will make the influence of the incentives estimated in this paper less obvious. However, some interesting questions to explore in future analysis could be, for example, to study the influence of referee decisions on other types of incentives such as salaries or referee relegation and to extend this estimation to other sports competitions and other industries in which the outcome depends on subjective decisions.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S. (2015) TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. (Available from https://www.tensorflow.org).

Breiman, L. (2001). Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author). *Statistical Science* 16. Institute of Mathematical Statistics: 199–231.

Buraimo, B., Forrest, D. and Simmons, R. (2010) The twelfth man? Refereeing bias in English and German soccer. *Journal of the Royal Statisticial Society: Series A (Statistics in Society),* 173, 431-449.

Dawson, P., and Dobson, S. (2010) The influence of social pressure and nationality on individual decisions: Evidence from the behaviour of referees. *Journal of Economic Psychology*, 31, 181-191.

Dawson, P., Dobson, S., Goddard, J. and Wilson, J. (2007) Are football referees really biased and inconsistent? Evidence on the incidence of disciplinary sanction in the English Premier League. *Journal of the Royal Statisticial Society: Series A (Statistics in Society)*, 170, 231-250.

Duchi, J., Hazan, E. and Singer, Y. (2011) Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research,* 12, 2121–59.

Garicano, L., Palacios-Huerta, I. and Prendergast, C. (2005) Favoritism under social pressure. *Review of Economics and Statistics*, 87, 208-216.

Ioffe, S. and Szegedy, Ch. (2015) Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In ICML.

Pettersson-Lidbom, P. and Priks, M. (2010) Behavior under social pressure: Empty Italian stadiums and referee bias. *Economics Letters*, 108, 212-214.

Pope, D.G., Price, J. and Wolfers, J (2018) Awareness Reduces Racial Bias. *Management Science* (forthcoming)

Polson, N. G, Sokolov, V. (2017). Deep Learning: A Bayesian Perspective. *Bayesian Analysis* , 12, 1275–1304.

Price, J., Remer, M. and Stone, D.F. (2012) Subperfect game: Profitable biases and NBA preferences. *Journal of Economics and Management Strategy*, 21, 271-300.

Price, J. and Wolfers, J. (2010) Ratial discrimination among NBA players. *Quarterly Journal of Economics*, 125, 1859-1887.

Samuelson, W. and Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, 1, 7-59.

Schmidhuber, J. (2015). Deep Learning in Neural Networks: An Overview. *Neural Networks,* 61, 85–117.

Silberzahn, R., and Uhlmann, E. L. (2015) Many hands make tight work. *Nature*, 526, 189-191.

Stecklov, G., Weinreb, A., and Carletto, C. (2018) Can incentives improve survey data quality in developing countries?: results from a field experiment in India. *Journal of the Royal Statisticial Society, Series A (Statistics in Society)*, 181, 1033-1056.

Titiunik, R., and Feher, A. (2017) Legislative behaviour absent re-election incentives: findings from a natural experiment in the Arkansas Senate. *Journal of the Royal Statisticial Society: Series A (Statistics in Society)*, 181, 351–378.

Tversky, A., and Kahneman, D. (1974) Judgment under uncertainty: Heuristics and biases. *Science*; 185,1124–1131.

Wood, S.N, Pya, N. and Säfken, B. (2016). Smoothing Parameter and Model Selection for General Smooth Models. *Journal of the American Statistical Association* ,111,1548–63.